Batelòc : cap a una basa informatisada de tèxtes occitans

Myriam Bras, Jean Thomas

BaTelÒc : cap a una basa informatizada de tèxtes occitans

Myriam Bras

Universitat de Tolosa, CLLE-ERSS CNRS & Universitat Tolosa Lo Miralh,
myriam.bras@univ-tlse2.fr

Joan Thomas

Universitat de Tolosa, IUFM , joanthomas@club-internet.fr

1. Fons de donadas textualas pels estudis occitans

Una de las grandas dificultats del trabalh scientific dins lo domèni occitan es la del manca d'apleches per la recèrca. Per mai d'un cercaire, que siá linguista, literari, istorian, etnològ, sociològ, lo trabalh de recèrca passa per l'estudi de tèxtes en lenga occitana. Mas an pas totes l'astre de poder recampar aisidament lo fons de donadas necite, e lor cal plan sovent picar o numerisar los tèxtes que vòlon estudiar per tal d'aprestar lor fons de donadas.

Los cercaires mai astrucs, d'aquel punt de vista, son los cercaires medievistas que pòdon utilizar lo *Concordancièr de l'Occitan Medieval* (Rickets, 2001, 2005) e lo *Corpus Electronic de l'Occitan Gascon* (Field, 2008), dos corpus de tèxtes (literaris pel primièr, juridics pel segond) informatizats que permeton als cercaires de far de cèrcas eficaças e aviadas.

Pels tèxtes de las temporadas modèrna (sègles XVI-XVII) e contemporanèa (sègles XVIII-, XXI), se los libres e manescriches de papièr existisson, totes son pas de bon trapar, mai que mai pels cercaires que son endefòra d'Occitània. Mercés al trabalh de numerizacion menat per l'associacion *CIEL d'ÒC*, una bona tièra d'òbras (mai de 500 tèxtes dels sègles XIX e XX) es ara a posita dins la *Bibliotèca Virtuala de la Tor Manha* jos la forma de documents electronics encodats dins un format estandard, lo format pdf. Aquesta bibliotèca numerica es "en linha" [1], es a dire consultabla per totòm via internet. La màger part dels tèxtes son

[1] http://sites.univ-provence.fr/tresoc/

numerizats en mòde tèxte, çò que permet de ne copiar de tròças per bastir lo fons de donadas necites a un trabalh de recèrca. Mas per aquò far, cal causir los tèxtes los uns après los autres, anar quèrre dins cada tèxte lo tròç causit e bastir "a la man" (en fasent de "talhatges/pegatges") lo fons de donadas.

2. Cap a una Basa Textuala per l'occitan : BaTelÒc

Per ajudar la recèrca a partir dels tèxtes occitans, una bibliotèca numerica coma la *Bibliotèca Virtuala de la Tor Manha* constituís ja una ressorga preciosa. Mas se pòt anar mai luènh per ajudar los cercaires al moment que vòlon bastir lor fons de donadas amb un dispositiu que se ditz una « basa textuala ». Una basa textuala recampa, coma una bibliotèca electronica, de tèxtes numerizats. Mas i a doas diferenças màgers.

La primièra es qu'una basa textuala es una « basa de donadas » que sas donadas son de tèxtes. Es doncas un ensem organisat e estructurat de tèxtes que pòdon èsser manejats – classats, triats…– segond una tièra de critèris. Aqueles critèris son d'informacion suls tèxtes que se dison las « metadonadas », per exemple lo nom de l'autor dels tèxtes a seleccionar, lor data o temporada d'escritura, lo dialècte, etc. Aquò permet de considerar pas totes los tèxtes de la basa al còp, mas daissa a cadun la possibilitat de se bastir son « corpus de trabalh » en causissent dins la basa los tèxtes que vòl estudiar. Una basa textuala es pas un corpus del contengut fixe, mas una mena de « servador a corpus » (Habert, 2000) ont cada cercaire pòt anar posar çò que li cal per bastir son fons de donadas, son « corpus de trabalh ».

La segonda diferença entre una bibliotèca numerica e una basa textuala es dins lo format e l'encodatge dels fichièrs. Los tèxtes d'una basa textuala son encodats dins un format que permet de far aisidament de cèrcas d'occurenças dins los tèxtes e que permet d'enriquesir lo tèxte amb d'annotacions divèrsas per fin de melhorar encara las cèrcas.

De mai en mai de lengas an lors basas textualas : *Frantext, The British National Corpus*, *Base de Datos Sintácticos del español actual, El Corpus Textual Informatitzat de la Llengua Catalana, XX. mendeko euskararen corpus estatistikoa, etc.* Aquelas basas son de ressorgas indispensablas per

menar de descripcions scientificas de las lengas dins los domènis del lexic, de la morfologia, de la sintaxi, etc. Mas las amiras de la bastison d'una basa textuala pòdon ésser tanben didacticas o lexicograficas (Bras e Thomas, 2007).

Lo projècte TelÒc (Bras, 2006) a per tòca de bastir una Basa Textuala per l'Occitan – *BaTelÒc* – en recampar d'òbras escrichas de totas menas – literatura, teatre, conte, tèxtes tècnics, jornalistics, cronicas – de las temporadas modèrna e contemporanèa. La basa textuala *BaTelÒc* e la *Bibliotèca Virtuala de la Tor Manha* seràn complementàrias, dins la mesura que metran pas los tèxtes a posita del meteis biais. Se pensa tanben dins un autre temps i apondre de tèxtes d'oral-escrich. La bastison d'una basa de tèxtes orals es ja aviada dins l'encastre del projècte del Thesaurus Occitan, *THESOC*[2]. S'agís d'enregistraments sonòrs que son tanben transcriches. *BaTelÒc* poiriá èsser interfaçada amb lo modul tèxte del *THESOC* per çò que concernís l'oral-escrich coma los contes e lo teatre.

En mai dels ligams amb aqueles dos projèctes occitans, la bastison de *BaTelÒc* profecharà d'un partenariat amb l'ATILF a Nancy, lo laboratòri que manten la basa textuala francesa *Frantext*[3]. *BaTelÒc* serà mesa a posita del public per nòstre labòratori CLLE-ERSS sus un portanèl dedicat a las ressorgas lexicalas e textualas manejat pel Centre Nacional de Ressorgas Textualas e Lexicalas (CNRTL)[4].

3. La basa experimentala

La primièra etapa del projècte es la bastison d'una basa experimentala d'un milion de mots. Per aquò far, mercés a l'ajuda de Robèrt Martí d'IDECO, avèm recampat un trentenat d'òbras recentas (doncas ja al format numeric) causidas dins la literatura de Provença, Gasconha, Lengadòc, Lemosin e Auvèrnha. Avèm encodats los tèxtes, dins lo lengatge XML (*eXtensible Markup Language*) en seguissent las nòrmas internacionalas d'encodatge de tèxtes de la TEI (*Text Encoding Initiative*) dichas P5. Lo prototip de motor de cèrca es ara operacional e la mesa en linha de la basa experimentala es prevista per 2009. Per ara lo trabalh es menat per

[2] http://thesaurus.unice.fr/
[3] http://www.frantext.fr/ e http://www.atilf.fr

una pichòta còla, avèm aguda l'ajuda tecnica de Marie-Paule Jacques puèi de Mai Ho-Dac[5] e de Franck Sajous pel codatge dels tèxtes e la programacion del motor de cèrca.

3.1 Codatge dels tèxtes

XML (eXtensible Markup Language) es un metalengatge de descripcion dels documents. Un document al format XML conten de "gavitèls" que marcan son estructura e las caractèristicas de sas partidas. La TEI (Text Encoding Initiative) es una nòrma generala per la descripcion dels tèxtes en scienças umanas. Los documents son estructurats coma un parelh Cap/Còs. Lo cap content las metadonadas, e lo còs conten lo tèxte. Donam çai-jós un exemple de tèxte encodat en XML a la nòrma TEI/P5.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!--<!DOCTYPE TEI SYSTEM "TeiP5.dtd">-->
<TEI>
        <teiHeader>
                                        ….. Un centenat de linhas de metadonadas ……
         </teiHeader>
        <text>
                        <front>
                                <doctitle><hi rend="M">l'estilò negre de la pluma d'aur</hi>
</doctitle>
                        </front>
                        <body>
                                <div type="preface">
                                <p>Sèm plan astrucs ! Òc ben, plan astrucs ! Ai l'estilò negre de la
pluma d'aur, l'estilò de totas las istòrias, de totas las jòias e de totas las lagremas, de tots los espèrs e
de totas las ràbias. </p
                                        ….. Aqui i a tot lo tèxte ondrat de gavitèls ……
                        </body>
        </text>
</TEI>
```

Per encodar lo tèxte segon aquel format, avèm mes al ponch una cadena de tractament : los fichièrs son pre-tractacts a la man abans d'èsser revirats automaticament en xml per un porgrama de traduccion ; las meta-donadas son sasidas mercés a una interfàcia especifica puèi los cap de fichièrs son generats automaticament.

---

[4] http://www.cnrtl.fr

[5] Mercés al sosten financièr del CNRTL e del CROM (Centre de Ressorgas Occitanas e Miegjornalas).

3.2 « Interrogar » BaTelòc

3.2.1 Causir lo corpus de trabalh

Tota utilizacion de Batelòc comença per la causida del corpus de trabalh, valent a dire tota la basa o una partida de la basa. La seleccion se fa en combinant de critèris coma lo nom de l'autor, lo titol de l'òbra, lo genre textual, lo dialècte de l'occitan, la nòrma grafica, la data o la temporada de publicación. Vesèm en Figura 1 un exemple de selección de totes los romans en lengadocian escriches en grafia classica de la basa experimentala amb lo prototip de motor de cèrca.

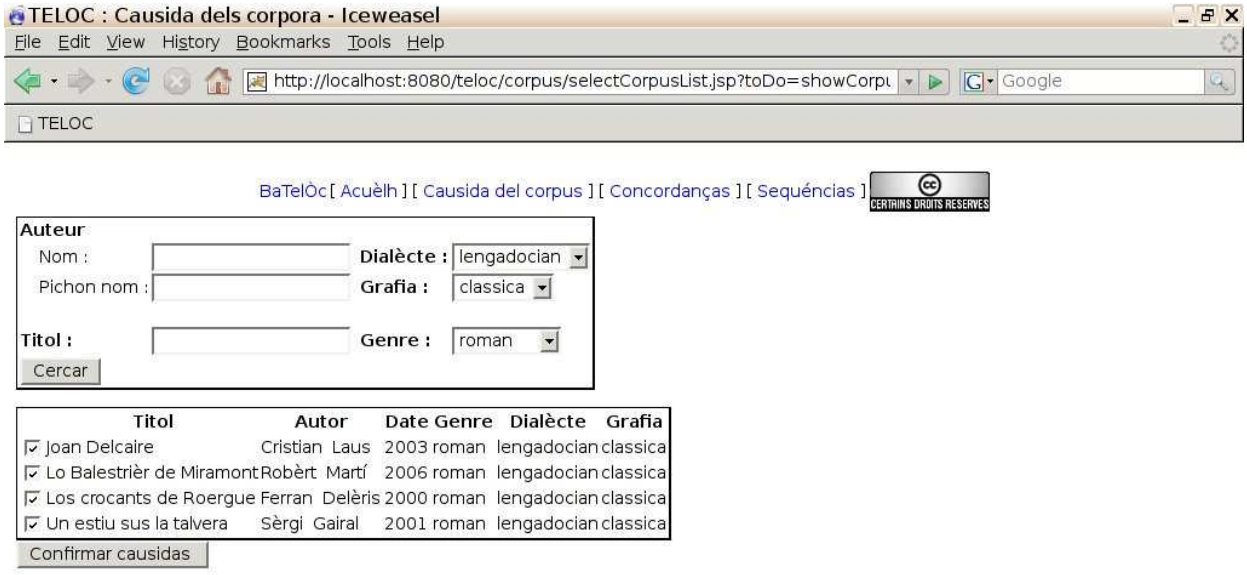

Figura 1 : Causida del Corpus de Trabalh

3.2.2 Cercar dins la basa

Un còp lo corpus de trabalh causit, se pòt far de requèstas de diferentas menas que permetan de cercar de concordanças, de tirar de contèxtes que clauson un mot, una partida o una sequéncia de mots, de cercar de coocurréncias. Es previst tanben de metre en òbra de calculs de frequéncias dins lo corpus.

La cèrca es basada sus la nocion de « forma ». Una « forma » pòt èsser un mot, una partida de mot un grop de mots (o per exemple), una frasa, etc.

Cèrca de concordanças d'una forma

La cèrca de concordanças d'una forma permet per exemple de far la tièra de totes los mots que s'acaban pel sufixe –*atièr*, de cercar lo « grop de mots » *qu'es aquò* (cf . Figura 2), o l'expression temporala *al cap d'una mesada* en cap de frasa per exemple (en utilizant la ponctuacion dins la requesta).

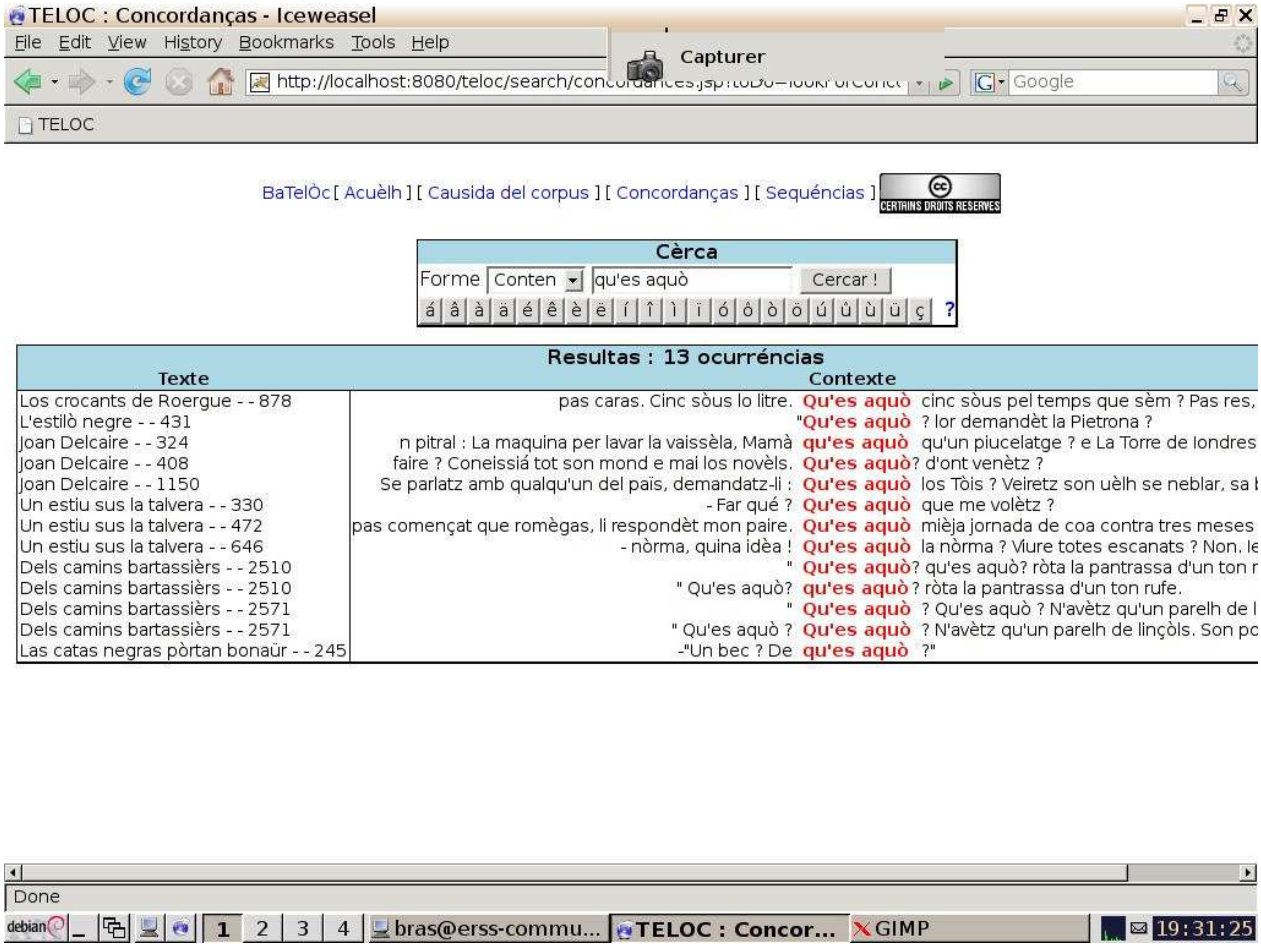

Figura 2 : Cèrca de las concordanças de la forma *qu'es aquò*

Cèrca de formas en contèxte

Aquela mena de cèrca es la meteissa que la de las concordanças mas la presentacion de las resultas n'es diferenta, coma o vesèm en Figura 3 que porgís las cèrcas dels contèxtes de l'expression *aquò rai*. Los contèxtes son mai largs que los de las concordanças.

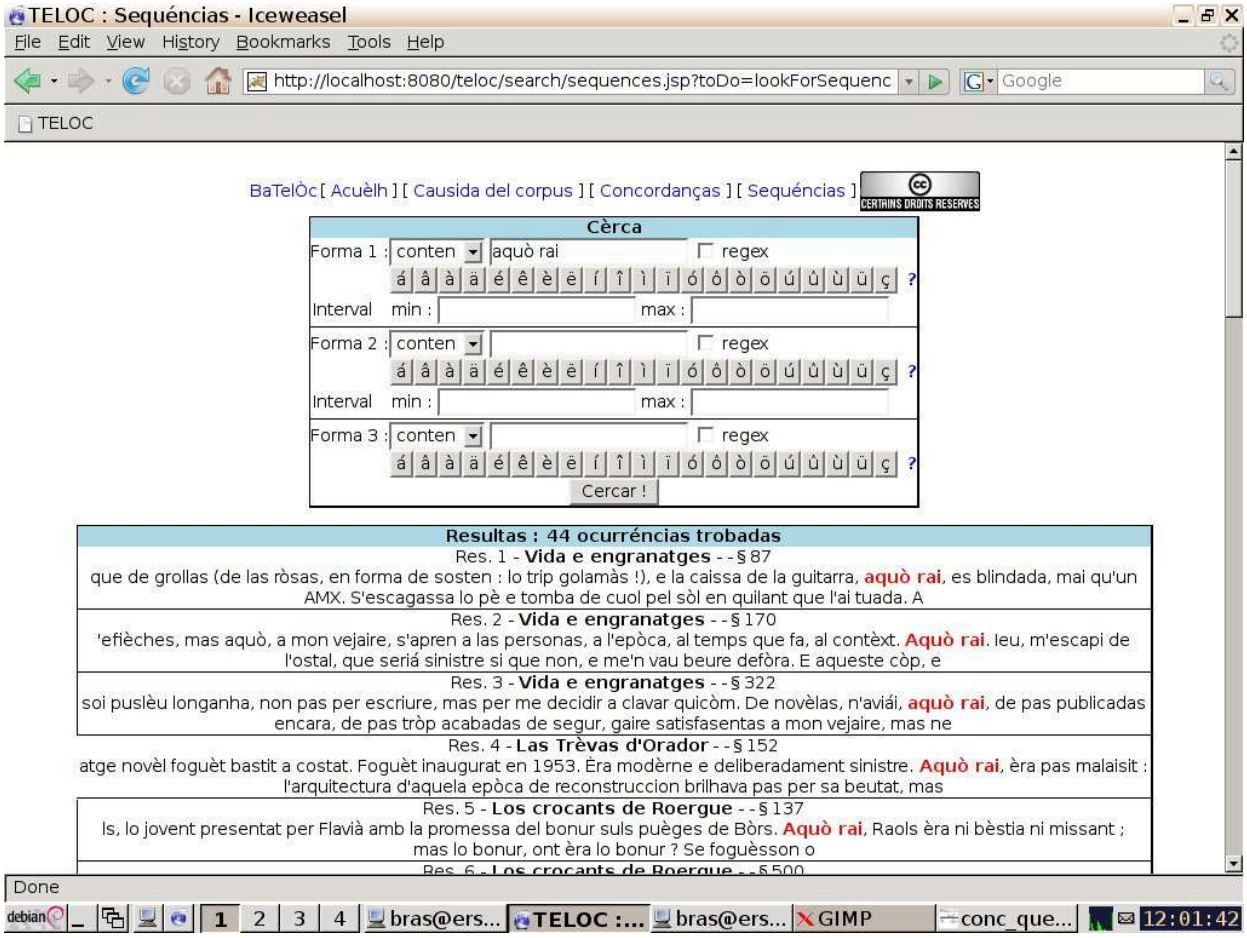

Figura 3 : Cèrca dels contèxtes de la forma *aquò rai*

Cèrca de sequéncias de formas

Una autra mena de cèrca consistís a traire los contèxtes que clauson de sequéncias de formas distantas dins lo tèxte. La figura 4 mostra per exemple que la cèrca de la séquencia *un costat … autre* dona de parelhs d'expressions coma :

d'un costat … de l'autre, d'un costat a l'autre, d'un costat… d'un autre…

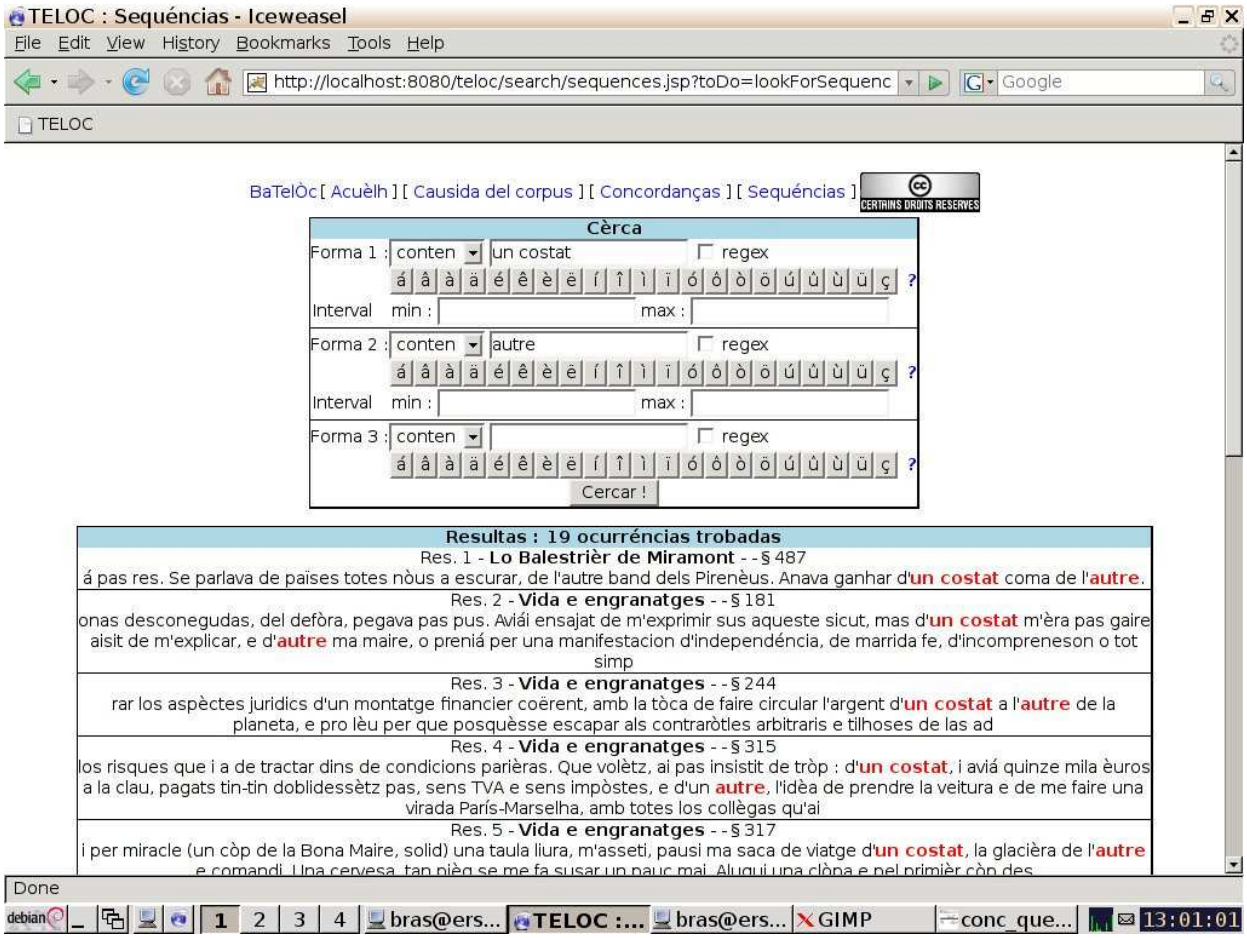

Figura 4 : Cèrca dels contèxtes de las sequéncias de formas *un costat….autre*

Cèrcas mai complèxas

Per fin de menar de cèrcas mai complèxes, se pòt utilizar un lengatge d'expressions regularas que permet de parametrizar las formas per cercar mai de causas al còp o de botar mai de constrenchas sus las resquestas.

Per exemple, se pòt cercar una forma o una autra forma dins la meteissa requesta, la disjonccion « o » es representada pel simbòl | : atal la requesta *solelh (colc|levat)* permet de cercar al còp la formas *solelh colc* e *solelh levat*. La requesta *qu'(es|ei) (aquò|çò)* permet de cercar *qu'es aquò, qu'es çò, qu'ei çò, qu'ei aquò.*

Lo lengatge d'expressions regularas permet tanben de cercar una forma en fin de mot. Per exemple, per cercar las occurréncias del sufixe *-et* , basta de far la requesta *\p{L}*et*. S'òm vòl pas fòrabandir los mots al plural, cal enriquesir la requesta en *\p{L}*(et|ets)* que va donar totes los mots que s'acaban per *et* ou *ets* : vèire en Figura 5 la cèrca dels contèxtes de mots que començan per *ram* et s'acaban per *et* o *ets*.

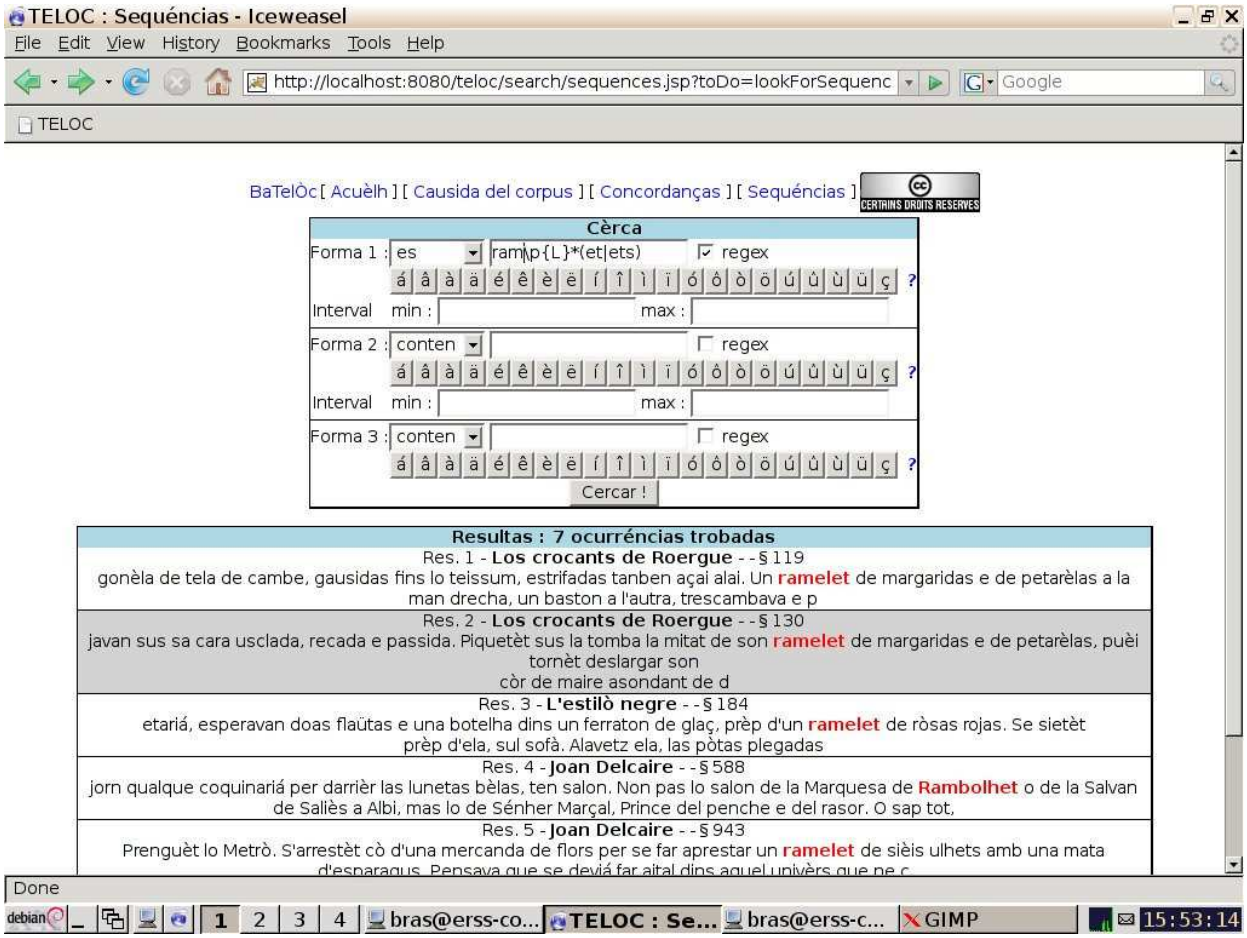

Figura 5 : Cèrca, amb l'ajuda de las expressions regularas (regex), dels mots que començan per *ram* e s'acaban per *et* o *ets*, e dels contèxtes d'aqueles mots

Se pòt de segur combinar aquelas diferentas menas de cèrcas : cèrcas de sequéncias de formas recuperadas amb d'expressions regularas, per exemple.

4. Cap a la basa operacionala

4.1 Far créisser la basa

Un còp la basa experimentala en plaça, aurem per tòca de far créisser lo nombre de tèxtes per fin de passar a una basa operacionala de qualque detzenats de milions de mots[6]. Dins aquela etapa, la *Bibliotèca Virtuala de la Tor Manha* poirà constituir una ressorga per *BaTelÒc* : los tèxtes ja numerisats per Ciel d'òc poiràn seguir la cadena de tractament necite a l'encodatge al format XML/TEI P5 per dintrar dins *BaTelÒc*. Dins l'autre sens, un utilizator de la *Bibliotèca Virtuala* poirà tanplan far virar los espleches d'interrogacion de *BaTelòc* per anar mai luènh dins l'estudi d'un tèxte que i aurà legit.

Coma la basa experimentala, la basa operacionala serà estructurada per genre, per temporada e data, per dialecte, per grafia, e per autor, mercés a las metadonadas estacadas a cada tèxte. Mas, per que se pòsca bastir de corpora de trabalh segon las règlas de bastison d'un corpus vertadièr – que los genres, domènis, dialectes e tipes de supòrt i devon èsser representats d'un biais equilibrat – calrà far créisser la basa en apondent de tèxtes de totas menas. Volèm seguir per aquò, non pas l'exemple de Frantext, que claus mai que mai de literatura, mas los exemples dels corpora bèlses de la lenga englesa The Bank of English[7] o The British National Corpus[8], e, mai generalament, aprofechar las experiençàs de constitucion de corpus representatius e equilibrats per d'autras lengas (veire Péry-Woodley, 1995 ; Habert, 2005). En mai de la literatura, la basa deurà doncas enclaure d'autras menas de tèxtes : de tèxtes d'oral-escrich coma los discors politics, los sermons, de tèxtes coma las cronicas e los articles d'almanacs o de revistas, de tèxtes tecnics, juridics o ligats a de mestièrs, las correspondéncias, emai las correspondéncias electronicas mai recentas, e mantunas menas de tèxtes orals transcriches.

---

[6] Per dire de se donar d'òrdres de pagèla, la basa FRANTEXT conten a l'entorn de 4000 tèxtes per 220 milions de mots, recampats en un quarantenat d'ans.

[7] http://www.cobuild.collins.co.uk

[8] http://www.natcorp.ox.ac.uk/

Los dreches d'autors seràn respectats coma se deu. Los tèxtes de la basa seràn, d'un biais general, pas accessibles al public dins lor integralitat. Coma o avèm vist amb la demonstracion de la basa experimentala, seràn pas que de colleccions de tròces de tèxtes coma resultas de las requèstas que seràn porgits. Per las òbras jos drech, los tròces de tèxtes faràn pas mai de 300 caractèrs. Per las òbras liuras de dreches, los tròces de tèxtes porgits poiran èsser mai largs, e l'accès a l'òbra integrala poirà èsser permesa, se ne vira, coma dins la bibliotèca numerica de Ciel d'Òc.

4.2 Annotar los tèxtes

Quand la basa serà pro larga (4 o 5 milions de mots), poirem aviar un tractament linguistic per etiquetar las formas graficas amb d'informacions morfosintacticas per ne far una basa lematizada e categorizada. Sus la basa enriquesida e annotada d'aquel biais, se poirà far de requèstas mai complèxas : cercar totas las formas conjugadas d'un vèrbe, totas las formas d'un adjectiu amb sas marcas de genre e de nombre ; téner compte de las categorias sintacticas dins las requèstas per cercar per exemple totas las prepausicions seguidas per un vèrbe ; combinar las cèrcas de co-occurréncia e amb las flexions verbalas, per cercar per exemple una frasa al preterit que comença per l'advèrbe *puèi* e que seguís una autra frasa al preterit. Se poirà tanben bastir de tièra de mots per enriquesir encara las requestas, far servir de gramaticas formalas per bastir de règlas parametrablas, o far de calculs de frequéncias, e d'estatisticas. Aquelas aisinas permetràn de tirar de jos-lexics o de glossaris, de tirar de collocacions, de cercar las primièras atestacions (dins la basa) d'un mot ou d'una forma. L'interés d'aquestas requestas complèxas per d'estudis linguistics, que sián morfologics, sintactics, lexicals o semantics, es evident. Mas d'autras disciplinas coma l'estilistica, la literatura, la glossaristica, la lexicografia, l'istòria, l'etnologia e la didactica de la lenga ne poiràn tanben aprofechar.

Lo trabalh d'annotacion necite es pas dels mendres, estant la manca de ressorgas lexicalas informatizadas e la variacion grafica, morfologica e lexicala de l'occitan. Mas poirem beneficiar per començar de l'etiquetaire e del lematizaire del projècte APERTIUM qu'a per tòca de desvolopar de traductors automatics occitan-catalan e occitan-espanhòl (Armentano i Oller, 2008).

5. Un projècte cooperatiu

La bastison de *BaTelÒc* es menada dins l'encastre d'un laboratòri de recèrca de linguistica (CLLE-ERSS, Universitat de Tolosa lo Miralh e CNRS) qu'a tanben de competenças en tractament automatic de las lengas, crucialas per nòstre prètzfach. De partenariats, crucials eles tanben, son ja en plaça amb IDECO/IEO, CIEL d'ÒC, lo CROM, l'ATILF e lo CNRTL. Aquel projècte es d'interés colectiu pels estudis occitans. Nos sembla que capitarà pas que se ven un projècte cooperatiu. Se pòt cooperar a dos nivèls al mens : cadun pòt èsser *utilizator,* es a dire utilizar *BaTelÒc* per sas recèrcas e nos far passar las criticas e idèas per melhorar la basa, son contengut e las aisinas de cèrca ; cadun pòt venir *contributor*, es a dire nos fisar un tèxte (format rtf, doc, txt ….) que vendrà enriquesir la basa. Convidam doncas totes los que son implicats dins la produccion de tèxtes occitans coma autor o editor, totes los qu'an de tèxtes occitans (escriches o orals) coma matèria de trabalh o objècte d'estudis a cooperar amb nautres per far créisser *BaTelÒc*[9].

Referenças bibliograficas

ARMENTANO i OLLER, C. (2008). Traduction automatique occitan-catalan et occitan-espagnol : difficultés affrontées et résultats atteints. In Neuvième Congrès International de l'Association Internationale d'Etudes Occitanes, Aachen. AIEO.

BRAS, M. (2006). Le projet TelÒc : construction d'une base textuelle occitane. Langues et Cité : bulletin de l'observation des pratiques linguistiques, 8:9.

BRAS, M. et THOMAS, J. (2007). Diccionaris, corpora, e basas de donadas textualas. Linguistica Occitana, 5:1–22.

FIELD, T. (2008). Langue et société au XIII^{ème} siècle à la lumière du corpus électronique du

[9] Dins aquela amira nòstre projècte poiriá cooperar amb lo de Gérard Ligozat de desvolopar un sistème de tractament automatic de tèxtes parallèls (Ligozat, 2008).

gascon médiéval. In Neuvième Congrès International de l'Association Internationale d'Etudes Occitanes, Aachen. AIEO.

HABERT, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In BILGER, M., éditeur : Linguistique sur corpus. Etudes et réflexions, numéro 31 de Cahiers de l'université de Perpignan, pages 11–58. Presses Universitaires de Perpignan, Perpignan.

HABERT, B. (2005). Instruments et ressources électroniques pour le français. L'essentiel Français. Ophrys, Gap/Paris.

LIGOZAT, G. (2008). Traitement automatique de textes parallèles : le cas de l'occitan moderne. In Neuvième Congrès International de l'Association Internationale d'Etudes Occitanes, Aachen. AIEO.

PÉRY-WOODLEY, M.-P. (1995). Quels corpus pour quels traitements automatiques ? TAL, 36(1-2):213–232.

RICKETS, P. (2001). COM1 Concordance de l'Occitan Médiéval (CD-ROM). Brepols, Turnhout.

RICKETS, P. (2005). COM 2. Les Troubadours. Les Textes Narratifs en vers (CD-ROM). Brepols, Turnhout.